A Grain of Salt for the WMT Manual Evaluation*

Ondřej Bojar, Miloš Ercegovčević, Martin Popel
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{bojar,popel}@ufal.mff.cuni.cz
ercegovcevic@hotmail.com

Omar F. Zaidan
Department of Computer Science
Johns Hopkins University
ozaidan@cs.jhu.edu

Abstract

The Workshop on Statistical Machine Translation (WMT) has become one of ACL's flagship workshops, held annually since 2006. In addition to soliciting papers from the research community, WMT also features a shared translation task for evaluating MT systems. This shared task is notable for having *manual evaluation* as its cornerstone. The Workshop's overview paper, playing a descriptive and administrative role, reports the main results of the evaluation without delving deep into analyzing those results. The aim of this paper is to investigate and explain some interesting idiosyncrasies in the reported results, which only become apparent when performing a more thorough analysis of the collected annotations. Our analysis sheds some light on how the reported results should (and should not) be interpreted, and also gives rise to some helpful recommendation for the organizers of WMT.

1 Introduction

The Workshop on Statistical Machine Translation (WMT) has become an annual feast for MT researchers. Of particular interest is WMT's shared translation task, featuring a component for manual evaluation of MT systems. The friendly competition is a source of inspiration for participating teams, and the yearly overview paper (Callison-Burch et al., 2010) provides a concise report of the state of the art. However, the amount of interesting data collected every year (the system outputs and, most importantly, the annotator judgments) is quite large, exceeding what the WMT overview paper can afford to analyze with much depth.

In this paper, we take a closer look at the data collected in last year's workshop, WMT10[1], and delve a bit deeper into analyzing the manual judgments. We focus mainly on the English-to-Czech task, as it included a diverse portfolio of MT systems, was a heavily judged language pair, and also illustrates interesting "contradictions" in the results. We try to explain such points of interest, and analyze what we believe to be the positive and negative aspects of the currently established evaluation procedure of WMT.

Section 2 examines the primary style of manual evaluation: system ranking. We discuss how the interpretation of collected judgments, the computation of annotator agreement, and document that annotators' individual preferences may render two systems effectively incomparable. Section 3 is devoted to the impact of embedding reference translations, while Section 4 and Section 5 discuss some idiosyncrasies of other WMT shared tasks and manual evaluation in general.

2 The *System Ranking* Task

At the core of the WMT manual evaluation is the system ranking task. In this task, the annotator is presented with a source sentence, a reference translation, and the outputs of five systems over that source sentence. The instructions are kept minimal: the annotator is to rank the presented translations from best to worst. Ties are allowed, but the scale provides five rank labels, allowing the annotator to give a total order if desired.

The five assigned rank labels are submitted at once, making the 5-tuple a unit of annotation. In the following, we will call this unit a *block*. The blocks differ from each other in the choice of the

* This work has been supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), P406/10/P259, MSM 0021620838, and DARPA GALE program under Contract No. HR0011-06-2-0001. We are grateful to our students, colleagues, and the three reviewers for various observations and suggestions.

[1] http://www.statmt.org/wmt10

Language Pair	Systems	Blocks	Labels	Comparisons	Ref $\geq$ others	Intra-annot. κ	Inter-annot. κ
German-English	26	1,050	5,231	10,424	0.965	0.607	0.492
English-German	19	1,407	6,866	13,694	0.976	0.560	0.512
Spanish-English	15	1,140	5,665	11,307	0.989	0.693	0.508
English-Spanish	17	519	2,591	5,174	0.935	0.696	0.594
French-English	25	837	4,156	8,294	0.981	0.722	0.452
English-French	20	801	3,993	7,962	0.917	0.636	0.449
Czech-English	13	543	2,691	5,375	0.976	0.700	0.504
English-Czech	18	1,395	6,803	13,538	0.959	0.620	0.444
Average	19	962	4,750	9,471	0.962	0.654	0.494

Table 1: Statistics on the collected rankings, quality of references and kappas across language pairs. In general, a block yields a set of five rank labels, which yields a set of $\binom{5}{2} = 10$ pairwise comparisons. Due to occasional omitted labels, the Comparisons/Blocks ratio is not exactly 10.

source sentence and the choice of the five systems being compared. A couple of tricks are introduced in the sampling of the source sentences, to ensure that a large enough number of judgments is repeated across different screens for meaningful computation of inter- and intra-annotator agreement. As for the sampling of systems, it is done uniformly – no effort is made to oversample or undersample a particular system (or a particular pair of systems together) at any point in time.

In terms of the interface, the evaluation utilizes the infrastructure of Amazon's Mechanical Turk (MTurk)[2], with each MTurk HIT[3] containing three blocks, corresponding to three consecutive source sentences.

Table 1 provides a brief comparison of the various language pairs in terms of number of MT systems compared (including the reference), number of blocks ranked, the number of pairwise comparisons extracted from the rankings (one block with 5 systems ranked gives 10 pairwise comparisons, but occasional unranked systems are excluded), the quality of the reference (the percentage of comparisons where the reference was better or equal than another system), and the κ statistic, which is a measure of agreement (see Section 2.2 for more details).[4]

We see that English-to-Czech, the language pair on which we focus, is not far from the average in all those characteristics except for the number of collected comparisons (and blocks), making it the second most evaluated language pair.

[2] `http://www.mturk.com/`

[3] "HIT" is an acronym for *human intelligence task*, which is the MTurk term for a single screen presented to the annotator.

[4] We only use the "expert" annotations of WMT10, ignoring the data collected from paid annotators on MTurk, since they were not part of the official evaluation.

2.1 Interpreting the Rank Labels

The description in the WMT overview paper says: "Relative ranking is our official evaluation metric. [Systems] are ranked based on how frequently they were judged to be **better than or equal to any other system**." (Emphasis added.) The WMT overview paper refers to this measure as "$\geq$ others", with a variant of it called "$>$ others" that does not reward ties.

We first note that this description is somewhat ambiguous, and an uninformed reader might interpret it in one of two different ways. For some system A, each block in which A appears includes four implicit pairwise comparisons (against the other presented systems). How is A's score computed from those comparisons?

The correct interpretation is that A is rewarded once for **each** of the four comparisons in which A wins (or ties).[5] In other words, A's score is the number of pairwise comparisons in which A wins (or ties), divided by the total number of pairwise comparisons involving A. We will use "$\geq$ others" (resp. "$>$ others") to refer to this interpretation, in keeping with the terminology of the overview paper.

The other interpretation is that A is rewarded only if A wins (or ties) **all** four comparisons. In other words, A's score is the number of *blocks* in which A wins (or ties) all comparisons, divided by the number of *blocks* in which A appears. We will use "$\geq$ all in block" (resp. "$>$ all in block") to refer to this interpretation.[6]

[5] Personal communication with WMT organizers.

[6] There is yet a third interpretation, due to a literal reading of the description, where A is rewarded at most once per block if it wins (or ties) *any one* of its four comparisons. This is probably less useful: it might be good at identifying the bottom tier of systems, but would fail to distinguish between all other systems.

	REF	CU-BOJAR	CU-TECTO	EUROTRANS	ONLINEB	PC-TRANS	UEDIN
$\geq$ others	95.9	65.6	60.1	54.0	**70.4**	62.1	62.2
$>$ others	90.5	45.0	44.1	39.3	49.1	**49.4**	39.6
$\geq$ all in block	93.1	32.3	30.7	23.4	**37.5**	32.5	28.1
$>$ all in block	81.3	13.6	**19.0**	13.3	15.6	18.7	10.6

Table 2: Sentence-level ranking scores for the WMT10 English-Czech language pair. The "$\geq$ others" and "$>$ others" scores reproduced here exactly match numbers published in the WMT10 overview paper. A boldfaced score marks the best system in a given row (besides the reference).

For quality control purposes, the WMT organizers embed the reference translations as a 'system' alongside the actual entries (the idea being that an annotator clicking randomly would be easy to detect, since they would not consistently rank the reference 'system' highly). This means that the reference is as likely as any other system to appear in a block, and when the score for a system A is computed, pairwise comparisons with the reference *are* included.

We use the publicly released human judgments[7] to compute the scores of systems participating in the English-Czech subtask, under both interpretations. Table 2 reports the scores, with our "$\geq$ others" (resp. "$>$ others") scores reproduced exactly matching those reported in Table 21 of the WMT overview paper. (For clarity, Table 2 is abbreviated to include only the top six systems of twelve.)

Our first suggestion is that **both** measures could be reported in future evaluations, since each tells us something different. The first interpretation gives partial credit for an MT system, hence distinguishing systems from each other at a finer level. This is especially important for a language pair with relatively few annotations, since "$\geq$ others" would produce a larger number of data points (four per system per block) than "$\geq$ all in block" (one per system per block). Another advantage of the official "$\geq$ others" is greater robustness towards various factors like the number of systems in the competition, the number of systems in one block or the presence of the reference in the block (however, see Section 3).

As for the second interpretation, it helps identify whether or not a single system (or a small group of systems) is strongly dominant over the other systems. For the systems listed in Table 2,

[7] http://statmt.org/wmt10/results.html

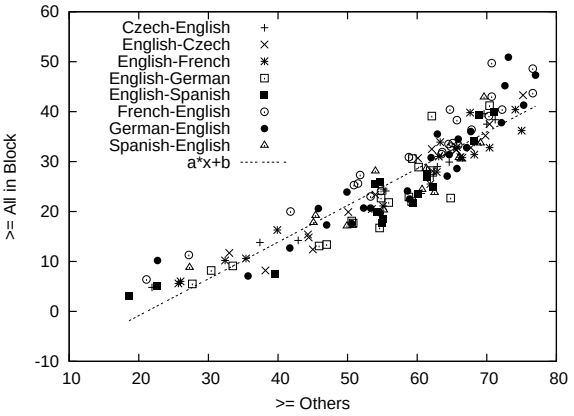

Figure 1: "$\geq$ all in block" and "$\geq$ others" provide very similar ordering of systems.

"$>$ all in block" suggests its potential in the context of system combination: CU-TECTO and PC-TRANS win almost one fifth of the blocks in which they appear, despite the fact that either a reference translation or a combination system already appears alongside them. (See also Table 4 below.)

Also, note that if the ranking task were designed specifically to cater to the "$\geq$ all in block" interpretation, it would only have **two** 'rank' labels (basically, "top" and "non-top"). In that case, annotators would spend *considerably* less time per block than they do now, since all they need to do is identify the top system(s) per block, without distinguishing non-top systems from each other.

Even for those interested in distinguishing non-state-of-the-art systems from each other, we point out that the "$\geq$ all in block" interpretation ultimately gives a system ordering that is very similar to that of the official "$\geq$ others" interpretation, **even** for the lower-tier systems (Figure 1).

2.2 Annotator Agreement

The WMT10 overview paper reports inter- and intra-annotator agreement over the pairwise comparisons, to show the validity of the evaluation setup and the "$\geq$ others" metric. Agreement is quantified using the following formula:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

where $P(A)$ is the proportion of times two annotators are observed to agree, and $P(E)$ is the expected proportion of times two annotators would agree by chance. Note that κ has a value of at most 1, with higher κ values indicating higher rates of agreement. The κ measure is more meaningful

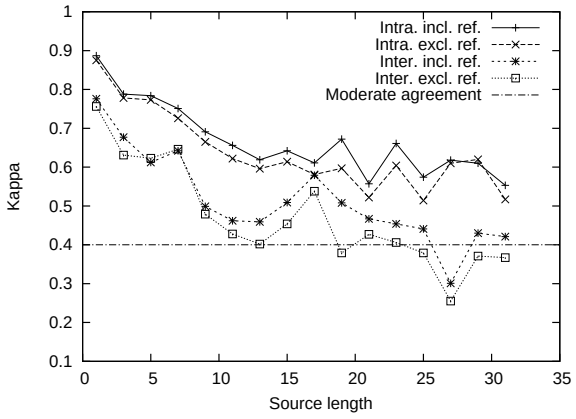

Figure 2: Intra-/inter-annotator agreement with/without references, across various source sentence lengths (lengths of n and $n+1$ are used to plot the point at $x = n$). This figure is based on all language pairs.

than reporting $P(A)$ as is, since it takes into account, via $P(E)$, how 'surprising' it is for annotators to agree in the first place.

In the context of pairwise comparisons, an agreement between two annotators occurs when they compare the same pair of systems (S_1,S_2), and both agree on their relative ranking: either $S_1 > S_2$, $S_1 = S_2$, or $S_1 < S_2$. $P(E)$ is then:

$$P(E) = P^2(_{S_1>S_2}) + P^2(_{S_1=S_2}) + P^2(_{S_1<S_2}) \quad (2)$$

In the WMT overview paper, all three categories are assumed equally likely, giving $P(E) = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}$. For consistency with the WMT overview paper, and unless otherwise noted, we also use $P(E) = \frac{1}{3}$ whenever a κ value is reported. (Though see Section 2.2.2 for a discussion about $P(E)$.)

2.2.1 Observed Agreement for Different Sentence Lengths

In Figure 2 we plot the κ values across different source sentence lengths. We see that the inter-annotator agreement (when excluding references) is reasonably high only for sentences up to 10 words in length – according to Landis and Koch (1977), and as cited by the WMT overview paper, not even 'moderate' agreement can be assumed if κ is less than 0.4. Another popular (and controversial) rule of thumb (Krippendorff, 1980) is more strict and says that $\kappa < 0.67$ is not suitable even for tentative conclusions.

For this reason, and given that a majority of sentences are indeed more than 10 words in length (the median is 20 words), we suggest that future evaluations either include fewer outputs per block, or divide longer sentences into shorter segments (e.g. on clause boundaries), so these segments are more easily and reliably comparable. The latter suggestions assumes word alignment as a preprocessing and presenting the annotators the context of the judged segment.

2.2.2 Estimating $P(E)$, the Expected Agreement by Chance

Several agreement measures (usually called kappas) were designed based on the Equation 1 (see Artstein and Poesio (2008) and Eugenio and Glass (2004) for an overview and a discussion). Those measures differ from each other in how to define the individual components of Equation 2, and hence differ in what the expected agreement by chance ($P(E)$) would be:[8]

- The S measure (Bennett et al., 1954) assumes a uniform distribution over the categories.
- Scott's π (Scott, 1955) estimates the distribution empirically from *actual annotation*.
- Cohen's κ (Cohen, 1960) estimates the distribution empirically as well, and further assumes *a separate distribution for each annotator*.

Given that the WMT10 overview paper assumes that the three categories ($S_1 > S_2$, $S_1 = S_2$, and $S_1 < S_2$) are equally likely, it is using the S measure version of Equation 1, though it does not explicitly say so – it simply calls it "the kappa coefficient" (K).

Regardless of what the measure should be called, we believe that the uniform distribution itself is not appropriate, even though it seems to model a "random clicker" adequately. In particular, and given the design of the ranking interface, $\frac{1}{3}$ is an overestimate of $P(S_1 = S_2)$ for a random clicker, and should in fact be $\frac{1}{5}$: each system receives one of five rank labels, and for two systems to receive the same rank label, there are only five (out of 25) label pairs that satisfy $S_1 = S_2$. Therefore, with $P(S_1 = S_2) = \frac{1}{5}$,

[8] These three measures were later generalized to more than two annotators (Fleiss, 1971; Bartko and Carpenter, 1976), Thus, without loss of generality, our examples involve two annotators.

"≥ Others"		S	π
Inter	incl. ref.	0.487	0.454
	excl. ref.	0.439	0.403
Intra	incl. ref.	0.633	0.609
	excl. ref.	0.601	0.575

Table 3: Summary of two variants of kappa: S (or K as it is reported in the WMT10 paper) and our proposed Scott's π. We report inter- vs. intra-annotator agreement and collected from all comparisons ("incl. ref.") vs. collected only from comparisons without the reference ("excl. ref.") because it is generally easier to agree that the reference is better than the other systems. This table is based on all language pairs.

we have $P(S_1 > S_2) = P(S_1 < S_2) = \frac{2}{5}$, and therefore $P(E) = 0.36$ rather than 0.333.

Taking the discussion a step further, we actually advocate following the idea of Scott's π, whereby the distribution of each category is estimated *empirically from the actual annotation*, rather than assuming a random annotator – these frequencies are easy to compute, and reflect a more meaningful $P(E)$.[9]

Under this interpretation, $P(S_1 = S_2)$ is calculated to be 0.168, reflecting the fraction of pairwise comparisons that correspond to a tie. (Note that this further supports the claim that setting $P(S_1 = S_2) = \frac{1}{3}$ for a random clicker, as used in the WMT overview paper, is an overestimate.) This results in $P(E) = 0.374$, yielding, for instance, $\pi = 0.454$ for "≥ others" inter-annotator agreement, somewhat lower than $\kappa = 0.487$ (reported in Table 3).

We do note that the difference is rather small, and that our aim is to be mathematically sound above all. Carefully defining $P(E)$ would be important when comparing kappas across different tasks with different $P(E)$, or when attempting to satisfy certain thresholds (as the cited 0.4 and 0.67). Furthermore, if one is interested in measuring agreement for individual annotators, such as identifying those who have unacceptably low intra-annotator agreement, the question of $P(E)$ is quite important, since annotation behavior varies noticeably from one annotator to another. A 'conservative' annotator who prefers to rank systems as being tied most of the time would have a high $P(E)$, whereas an annotator using ties moderately would have a low $P(E)$. Hence, two annotators with equal agreement rates ($P(A)$) are not necessarily equally proficient, since their $P(E)$ might differ considerably.[10]

2.3 The ≥ variant vs. the > variant

Even within the same interpretation of how systems could be scored, there is a question of whether or not to reward ties. The overview paper reports both variants of its measure, but does not note that there are non-trivial differences between the two orderings. Compare for example the "≥ others" ordering vs. the "> others" ordering of CU-BOJAR and PC-TRANS (Table 2), showing an unexpected swing of 7.9%:

	≥ others	> others
CU-BOJAR	**65.6**	45.0
PC-TRANS	62.1	**49.4**

CU-BOJAR seems better under the ≥ variant, but loses out when only strict wins are rewarded. Theoretically, this could be purely due to chance, but the total number of pairwise comparisons in "≥ others" is relatively large (about 1,500 pairwise comparisons for each system), and ought to cancel such effects.

A similar pattern could be seen under the "all in block" interpretation as well (e.g. for CU-TECTO and ONLINEB). Table 4 documents this effect by looking at how often a system is the sole winner of a block. Comparing PC-TRANS and CU-BOJAR again, we see that PC-TRANS is up there with CU-TECTO and DCU-COMBO as the most frequent sole winners, winning 71 blocks, whereas CU-BOJAR is the sole winner of only 53 blocks. This is in spite of the fact that PC-TRANS actually appeared in slightly fewer blocks than CU-BOJAR (385 vs. 401).

One possible explanation is that the two variants ("≥" and ">") measure two subtly different things about MT systems. Digging deeper into Table 2's values, we find that CU-BOJAR is tied with another system $65.6 - 45.0 = 20.4\%$ of the time, while PC-TRANS is tied with another system only $62.1 - 49.4 = 12.7\%$ of the time. So it seems that PC-TRANS's output is *noticeably different* from another system more frequently than CU-BOJAR, which reduces the number of times that annotators

[9] We believe that $P(E)$ should not reflect the chance that two *random* annotators would agree, but the chance that two **actual** annotators would agree *randomly*. The two sound subtly related but are actually quite different.

[10] Who's more impressive: a psychic who correctly predicts the result of a coin toss 50% of the time, or a psychic who correctly predicts the result of a *die roll* 50% of the time?

Blocks	Sole Winner
305	Reference
73	CU-TECTO
71	PC-TRANS
70	DCU-COMBO
57	RWTH-COMBO
54	ONLINEB
53	CU-BOJAR
46	EUROTRANS
41	UEDIN
41	UPV-COMBO
175	One of eight other systems
409	No sole winner
1395	Total English-to-Czech Blocks

Table 4: A breakdown of the 1,395 blocks for the English-Czech task, according to which system (if any) is the sole winner. On average, a system appears in 388 blocks.

mark PC-TRANS as tied with another system.[11] In that sense, the "$\geq$" ranking is hurting PC-TRANS, since it does not benefit from its small number of ties. On the other hand, the "$>$" variant would not reward CU-BOJAR for its large number of ties.

The "$\geq$ others" score may be artificially boosted if several very similar systems (and therefore likely to be "tied") take part in the evaluation.[12] One possible solution is to completely disregard ties and calculate the final score as $\frac{\text{wins}}{\text{wins+losses}}$. We recommend to use this score instead of "$\geq$ others" ($\frac{\text{wins+ties}}{\text{wins+ties+losses}}$) which is biased toward often tied systems, and "$>$ others" ($\frac{\text{wins}}{\text{wins+ties+losses}}$) which is biased toward systems with few ties.

2.4 Surprise? Does the Number of Evaluations Affect a System's Score?

When examining the system scores for the English-Czech task, we noticed a surprising pattern: it seemed that the more times a system is sampled to be judged, the lower its "$\geq$ others" score ("$\geq$ all in block" behaving similarly). A scatter plot of a system's score vs. the number of blocks in which it appears (Figure 3) makes the pattern obvious.

We immediately wondered if the pattern holds in other language pairs. We measured Pearson's correlation coefficient within each language pair, reported in Table 5. As it turns out, English-Czech happened to be the one language pair where the 'correlation' is strongest, with only English-Spanish also having a somewhat strong correlation. Overall, though, there is a consistent trend that can be seen across the language pairs. Could it really be the case that the more often a system is judged, the worse its score gets?

[11] Indeed, PC-TRANS is a commercial system (manually) tuned over a long period of time and based on resources very different from what other participants in WMT use.

[12] In the preliminary WMT11 results, this seems to happen to four Moses-like systems (UEDIN, CU-BOJAR, CU-MARECEK and CU-TAMCHYNA) which have better "$\geq$ others" score but worse "$>$ others" score than CU-TECTO.

Source	Target	Correlation of Block Count vs. "$\geq$ Others"
English	Czech	-0.558
English	Spanish	-0.434
Czech	English	-0.290
Spanish	English	-0.240
English	French	-0.227
English	German	-0.161
French	English	-0.024
German	English	0.146
Overall		-0.092

Table 5: Pearson's correlation between the number of blocks where a system was ranked and the system's "$\geq$ others" score. (The reference itself is not included among the considered systems).

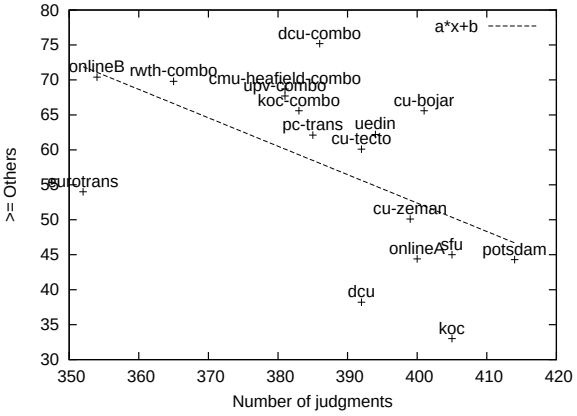

Figure 3: A plot of "$\geq$ others" system score vs. times judged, for English-Czech.

Examining plots for the other language pairs makes things a bit clearer. Consider for example the plot for English-Spanish (Figure 4). As one would hope, the data points actually come together to form a cloud, **indicating a lack of correlation**. The reason that a hint of a correlation exists is the presence of two outliers in the bottom right corner. In other words, the **very** worst systems are, indeed, the ones judged quite often. We observed this pattern in several other language pairs as well.

The correlation naturally does not imply causation. We are still not sure how to explain the artifact. A subtle possibility lies in the MTurk interface: annotators have the choice to accept a HIT or skip it before actually providing their la-

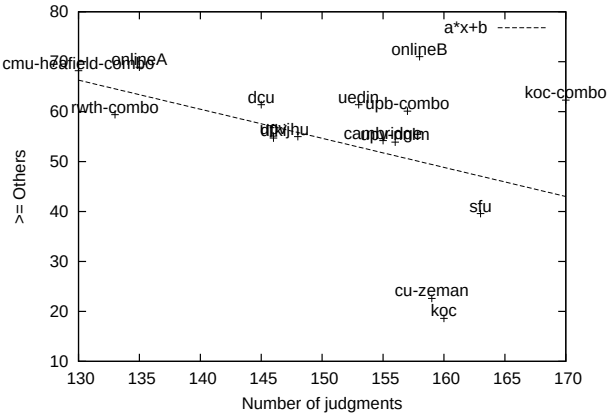

Figure 4: A plot of "≥ others" system score vs. times judged, for English-Spanish.

	REF	CU-BOJAR	CU-TECTO	EUROTRANS	ONLINEB	PC-TRANS	UEDIN
REF	-	4.3	4.3	5.1	3.8	3.6	2.3
CU-BOJAR	**87.1**	-	**45.7**	28.3	**44.4**	39.5	**41.1**
CU-TECTO	**88.2**	35.8	-	38.0	**55.8**	**44.0**	36.0
EUROTRANS	**88.5**	**60.9**	**46.8**	-	**50.7**	**53.8**	**48.6**
ONLINEB	**91.2**	31.1	29.1	32.8	-	43.8	**39.3**
PC-TRANS	**88.0**	**45.3**	42.9	28.6	**49.3**	-	36.6
UEDIN	**94.3**	39.3	**44.2**	31.9	32.1	**49.5**	-

Table 6: Pairwise comparisons extracted from sentence-level rankings of the WMT10 English-Czech News Task. Re-evaluated to reproduce the numbers published in WMT10 overview paper. Bold in column A and row B means that system A is pairwise better than system B.

bels. It might be the case that some annotators are more willing to accept HITs when there is an obviously poor system (since that would make their task somewhat easier), and who are more prone to skipping HITs where the systems seem hard to distinguish from each other. So there might be a causation effect after all, but in the reverse order: a system gets judged more often if it is a bad system.[13] A suggestion from the reviewers is to run a pilot annotation with deliberate inclusion of a poor system among the ranked ones.

2.5 Issues of Pairwise Judgments

The WMT overview paper also provides pairwise system comparisons: each cell in Table 6 indicates the percentage of pairwise comparisons between the two systems where the system in the column was ranked better (>) than the system in the row. For instance, there are 81 ranking responses where both CU-TECTO and CU-BOJAR were present and indeed ranked[14] among the 5 systems in the block. In 37 (45.7%) of the cases, CU-TECTO was ranked better, in 29 (35.8%), CU-BOJAR was ranked better and there was a tie in the remaining 15 (18.5%) cases. The ties are not explicitly shown in Table 6 but they are implied by the total of 100%. The cell is in bold where there was a win in the pairwise comparison, so 45.7 is bold in our example.

An interesting "discrepancy" in Table 6 is that CU-TECTO wins pairwise comparisons with CU-BOJAR and UEDIN but it scores worse than them in the official "≥ others", cf. Table 2. Similarly, UEDIN outperformed ONLINEB in the pairwise comparisons but it was ranked worse in both > and ≥ official comparison.

In the following, we focus on the CU-BOJAR (B) and CU-TECTO (T) pair because they are interesting competitors on their own. They both use the same parallel corpus for lexical mapping but operate very differently: CU-BOJAR is based on Moses while CU-TECTO transfers at a deep syntactic layer and generates target text which is more or less grammatically correct but suffers in lexical choice.

2.5.1 Different Set of Sentences

The mismatch in the outcomes of "≥ others" and pairwise comparisons could be caused by different set of sentences. The pairwise ranking is collected from the set of blocks where both CU-BOJAR and CU-TECTO appeared (and were indeed ranked). Each of the systems however competes in other blocks as well, which contributes to the official "≥ others".

The set of sentences underlying the comparison is very different and more importantly that the basis for pairwise comparisons is much smaller than the basis of the official "≥ others" interpretation. The outcome of the official interpretation however depends on the random set of systems your system was compared to. In our case, it is impossible to distinguish, whether CU-TECTO had just bad luck on sentences and systems it was compared to when CU-BOJAR was not in the block and/or whether the 81 blocks do not provide a reliable picture.

2.5.2 Pairwise Judgments Unreliable

To complement WMT10 rankings for the two systems and avoid the possible lower reliability due to 5-fold ranking instead of a targeted compari-

[13] No pun intended!

[14] The users sometimes did not fill any rank for a system. Such cases are ignored.

		Author of B says:				
		B>T	T>B	both fine	both wrong	Total
T says:	B>T	9	-	1	1	11
	T>B	2	13	-	3	18
	both fine	2	-	2	3	7
	both wrong	10	5	1	11	27
	Total	23	18	4	18	63

Table 7: Additional annotation of 63 CU-BOJAR (B) vs. CU-TECTO (T) sentences by two annotators.

	Better		Both	
Annotator	B	T	fine	wrong
A	**24**	23	5	11
C	10	**12**	5	36
D	**32**	20	2	9
M	11	**18**	7	27
O	**23**	18	4	18
Z	25	**27**	2	9
Total	**125**	118	25	110

Table 8: Blurry picture of pairwise rankings of CU-BOJAR vs. CU-TECTO. Wins in bold.

son, we asked the main authors of both CU-BOJAR and CU-TECTO to carry out a *blind* pairwise comparison on the exact set of 63 sentences appearing across the 81 blocks in which both systems were ranked. As the totals in Table 7 would suggest, each author unwittingly recognized his system and slightly preferred it. The details however reveal a subtler reason for the low agreement: one of the annotators was less picky about MT quality and accepted 10+5 sentences completely rejected by the other annotator. In total, these two annotators agreed on $9 + 13 + 2 + 11 = 35$ (56%) of cases and their pairwise κ is 0.387.

A further annotation of these 63 sentences by four more people completes the blurry picture: the pairwise κ for each pair of our five annotators ranges from 0.242 to 0.615 with the average 0.407±0.106. The multi-annotator κ (Fleiss, 1971) is 0.394 and all six annotators agree on a single label only in 24% of cases. The agreement is not better even if we merge the categories "Both fine" and "Both wrong" into a single one: The pairwise κ ranges from 0.212 to 0.620 with the average 0.405±0.116, the multi-annotator κ is 0.391. Individual annotations are given in Table 8.

Naturally, the set of these 63 sentences is not a representative sample. Even if one of the systems actually won, such an observation could not have been generalized to other test sets. The purpose of the exercise was to check whether we are *at all* able to agree which of the systems translates this specific set of sentences better. As it turns out, even a simple pairwise ranking can fail to provide an answer because different annotators simply have different preferences.

SRC	It's not completely ideal.		
REF	Není to úplně ideální.	Ranks	
PC-TRANS	To není úplně ideální.	2	5
CU-BOJAR	To není úplně ideální.	5	4

Table 9: Two rankings by the same annotator.

SRC	FCC awarded a tunnel in Slovenia for 64 million
REF	FCC byl přidělen tunel ve Slovinsku za 64 milionů
Gloss	FCC **was** awarded a tunnel in Slovenia for 64 million
HYP1	FCC přidělil tunel ve Slovinsku za 64 miliónů
HYP2	FCC přidělila tunel ve Slovinsku za 64 milionů
Gloss	FCC awarded$^{\text{masc}}_{\text{/fem}}$ a tunnel in Slovenia for 64 million

Figure 5: A poor reference translation confuses human judges. The SRC and REF differ in the active/passive form, attributing completely different roles to "FCC".

Finally, Table 9 illustrates how poor the WMT10 rankings can be. The exact same string produced by two systems was ranked differently each time – by the same annotator. (The hypothesis is a plausible translation, only the information structure of the sentence is slightly distorted so the translation may not fit well it the surrounding context.)

3 The Impact of the Reference Translation

3.1 Bad Reference Translations

Figure 5 illustrates the impact of poor reference translation on manual ranking as carried out in Section 2.5.2. Of our six independent annotations, three annotators marked the hypotheses as "both fine" given the match with the source and three annotators marked them as "both wrong" due to the mismatch with the reference. Given the construction of the WMT test set, this particular sentence comes from a Spanish original and it was most likely translated directly to both English and Czech.

Source	Target	Correlation of Reference vs. "≥ others"
Spanish	English	0.341
English	French	0.164
French	English	0.098
German	English	0.088
Czech	English	-0.041
English	Czech	-0.145
English	Spanish	-0.411
English	German	-0.433
Overall		-0.107

Table 10: Pearson's correlation of the relative percentage of blocks where the reference was included in the ranking and the final "≥ others" of the system (the reference itself is not included among the considered systems).

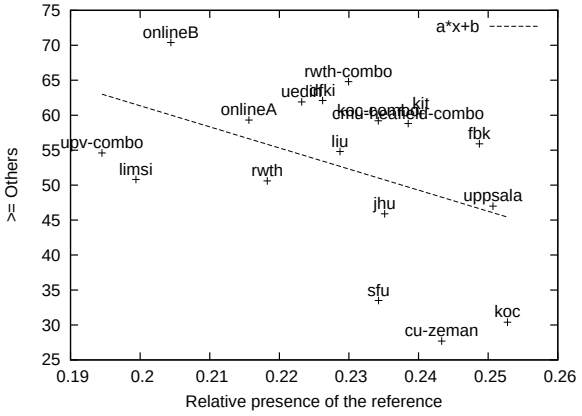

Figure 6: Correlation of the presence of the reference and the official "≥ others" for English-German evaluation.

3.2 Reference Can Skew Pairwise Comparisons

The exact set of competing systems in each 5-fold ranking in WMT10 evaluation is random. The "≥ others" however is affected by this: a system may suffer more losses if often compared to the reference, and similarly it may benefit from being compared to a poor competitor.

To check this, we calculate the correlation between the relative presence of the reference among the blocks where a system was judged and the system's official "≥ others" score. Across language, there is almost no correlation (Pearson's coefficient: -0.107). However, for some language pairs, the correlation is apparent, as listed in Table 10. Negative correlation means: the more often the system was compared along with the reference, the worse the score of the system.

Figure 6 plots the extreme case of English-German evaluation.

Source	Target	Min	Avg±StdDev	Max
English	Czech	40	65±19	115
English	French	40	66±17	110
English	German	10	40±16	80
English	Spanish	30	54±15	85
Czech	English	5	38±13	60
French	English	5	37±15	70
German	English	10	32±12	65
Spanish	English	35	56±11	70

Table 11: The number of post-edits per system for each language pair to complement Figure 3 (page 12) of the WMT10 overview paper.

4 Other WMT10 Tasks

4.1 Blind Post-Editing Unreliable

WMT often carries out one more type of manual evaluation: "Editing the output of systems without displaying the source or a reference translation, and then later judging whether edited translations were correct." (Callison-Burch et al., 2010). We call the evaluation "blind post-editing" for short.

We feel that blind post-editing is more informative than system ranking. First, it constitutes a unique comprehensibility test, and after all, MT should aim at comprehensible output in the first place. Second, blind post-editing can be further analyzed to search for specific errors in system output, see Bojar (2011) for a preliminary study.

Unfortunately, the amount of post-edits collected in WMT10 varied a lot across systems and language pairs. Table 11 provides the minimum, average and maximum number of post-edits of outputs of a particular MT system. We see that e.g. while English-to-Czech has many judgments of this kind per system, Czech-to-English is one of the worst supported directions.

It is not surprising that conclusions based on 5 observations can be extremely deceiving. For instance CU-BOJAR seems to produce 60% of outputs comprehensible (and thus wins in Figure 3 on page 12 in the WMT overview paper), far better than CMU. This is not in line with the ranking results where both rank equally (Table 5 on page 10 in the WMT overview paper). In fact, CU-BOJAR was post-edited 5 times and 3 of these post-edits were acceptable while CMU was post-edited 30 times and 5 of these post-edits were acceptable.

4.2 A Remark on System Combination Task

One results of WMT10 not observed in previous years was that system combinations indeed performed better than individual systems. Previous

	Dev Set	Test Set	
Sententes	455	2034	Diff
GOOGLE	17.32±1.25	16.76±0.60	↘
BOJAR	16.00±1.15	16.90±0.61	↗
TECTOMT	11.48±1.04	13.19±0.58	↗
PC-TRANS	10.24±0.92	10.84±0.46	↗
EUROTRAN	9.64±0.92	11.04±0.48	↗

Table 12: BLEU scores of sample five systems in English-to-Czech combination task.

years failed to show this clearly, because Google Translate used to be included among the combined systems, making it hard to improve. In WMT10, Google Translate was excluded from system combination task (except for translations involving Czech, where it was accidentally included).

Our Table 12 provides an additional explanation why the presence of Google among combined systems leads to inconclusive results. While the test set was easier (based on BLEU) than the development set for most systems, it was much harder for Google. All system combinations were thus likely to overfit and select Google n-grams most often. Without access to Google powerful language models, the combination systems were likely to underperform Google in final fluency of the output.

5 Further Issues of Manual Evaluation

We have already seen that the comprehensibility test by blind post-editing provides a different picture of the systems than the official ranking. Berka et al. (2011) introduced a third "quiz-based evaluation". The quiz-like evaluation used the English-to-Czech WMT10 systems, applied to different texts: short text snippets were translated and annotators were asked to answer three yes/no questions complementing each snippet. The order of the systems was rather different from the official WMT10 results: CU-TECTO won the quiz-based evaluation despite being the fourth in WMT10.

Because the texts were different in WMT10 and the quiz-based evaluation, we asked a small group of annotators to apply the ranking technique on the text snippets. While not exactly comparable to the WMT10 ranking, the WMT10 ranking was confirmed: CU-TECTO was again among the lowest-scoring systems and Google won the ranking.

Bojar (2011) applies the error-flagging manual evaluation by Vilar et al. (2006) to four systems of WMT09 English-to-Czech task. Again, the overall order of the systems is somewhat different when ranked by the number of errors flagged. Mireia Farrús and Fonollosa (2010) use a coarser but linguistically motivated error classification for Catalan-Spanish and suggest that differences in ranking are caused by annotators treating some types of errors as more serious.

In short, different types of manual evaluations lead to different results even when identical systems and texts are evaluated.

6 Conclusion

We took a deeper look at the results of the WMT10 manual evaluation, and based on our observations, we have some recommendations for future evaluations:

- We propose to use a score which ignores ties instead of the official "$\geq$ others" metric which rewards ties and "$>$ others" which penalizes ties. Another score, "$\geq$ all in block", could help identify which systems are more dominant.
- Inter-annotator agreement decreases dramatically with sentence length; we recommend including fewer sentences per block, at least for longer sentences.
- We suggest agreement be measured based on an empirical estimate of $P(E)$, or at least using a more correct random clicking $P(E) = 0.36$.
- There is evidence of a negative correlation between the number of times a system is judged and its score; we recommend a deeper analysis of this issue.
- We recommend the reference be sampled at a lower rate than other systems, so as to play a smaller role in the evaluation. We also recommend better quality control over the production of the references.

And to the readers of the WMT overview paper, we point out:

- Pairwise comparisons derived from 5-fold rankings are sometimes unreliable. Even a targeted pairwise comparison of two systems can shed little light as to which is superior.
- The acceptability of post-edits is sometimes very unreliable due to the low number of observations.

References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

John J. Bartko and William T. Carpenter. 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163(5):307–317.

E. M. Bennett, R. Alpert, and A. C. Goldstein. 1954. Communications through limited questioning. *Public Opinion Quarterly*, 18(3):303–308.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:77–86, March.

Ondřej Bojar. 2011. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA. Chapter 12.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

José B. Mariño Mireia Farrús, Marta R. Costa-jussà and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the Euoropean Association for Machine Translation (EAMT'10)*, pages 167–173, May.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May.